МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра биомедицинской инженерии

УТВЕРЖДАЮ
Проректор по учебной работе
_____ О.Г. Локтионова
«11» 09 2023 г.

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ КЛАССИФИКАЦИИ И РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ

Методические указания к лабораторным работам по дисциплине «Интеллектуальные системы классификации и распознавания изображений» для студентов направления подготовки 12.04.04 «Биотехнические системы и технологии»

Курск 2023

УДК 004.93:61

Составители: С.А. Филист.

Рецензент
Доктор технических наук, профессор Р.А. Томакова

Интеллектуальные системы классификации и распознавания изображений: методические указания к лабораторным работам / Юго-Зап. гос. ун-т; сост.: С.А. Филист. - Курск, 2023. - 33 с.

Предназначено для студентов по дисциплине «Интеллектуальные системы классификации и распознавания изображений» направления подготовки 12.04.04 «Биотехнические системы и технологии».

Текст печатается в авторской редакции

Подписано в печать . Формат 60×84 1/16. Бумага офсетная.
Усл. печ. л. 1,9. Уч.-изд. л. 1,7. Тираж 100 экз. Заказ 49 .
Юго-Западный государственный университет.
305040, г. Курск, ул. 50 лет Октября, 94.

ЛАБОРАТОРНАЯ РАБОТА №1. СОСТАВЛЕНИЕ ОБУЧАЮЩИХ ТАБЛИЦ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ И ИХ ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА

Задание на лабораторную работу

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 2

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).

4. Используя пакет STATISTICA определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 3

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по шесть вариационных рядов X1…X6 и X11…66. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 4

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку

входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.

2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 5

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.

9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 6

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по шесть вариационных рядов X1…X6 и X11…66. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 7

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.

5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 8

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 9

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по шесть вариационных рядов X1…X6 и X11…66.

Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.

2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 10

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 11

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 12

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по шесть вариационных рядов X1…X6 и X11…66. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.

7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 13

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 14

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с равномерным законом распределения.
2. По одномерным гистограммам оценить структуру классов.

3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 15

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по шесть вариационных рядов X1…X6 и X11…66. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 16

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с равномерным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Обычно в распоряжении исследователя имеются лишь данные выборки, например, значения количественного признака $x_1, x_2, \ldots x_n$, полученные в результате n наблюдений. Через эти данные и выражают оцениваемый параметр. При n > 50 для оценки математического ожидания и дисперсии следует пользоваться формулами 1.1 и 1.2 соответственно:

$$M[x] = \sum x_i / n, \qquad (1.1)$$

$$D[x] = \sum (x_i - M[x])^2 n, \qquad (1.2)$$

где n – длина выборки;

$$СКО = D[x]^{\frac{1}{2}} \qquad (1.3)$$

Медианой называется то значение, которое удовлетворяет условию:

$$P(x > M) = P(x < M), \qquad (1.4)$$

где M – медиана.

Модой называется то возможное значение, при которой плотность распределения максимальна.

Коэффициент вариации:

$$V = D[x]^{\frac{1}{2}} \cdot 100\% / M \qquad (1.5)$$

Коэффициент асимметрии:

$$As = (\sum (x_i - M)^3) / D[x]^{\frac{3}{2}}. \qquad (1.6)$$

Для статистического описания материала используются коэффициент корреляции, выборочный коэффициент отношения, которые вычисляются по формулам (3.3), (3.4) соответственно:

$$R_{xy} = \frac{\sum (x_i - M[x]) \cdot (y_i - M[y])}{(\sum (x_i - M[x])^2 \cdot \sum (y_i - M[y])^2)^{\frac{1}{2}}} \qquad (1.7)$$

$$\eta_{xy} = \delta_{yx} / \delta, \qquad (1.8)$$

где $\delta_{yx} = \sqrt{\sum n_x \cdot (y_{x\,ср} - y_{ср})^2 / n}$, $\delta_y = \sqrt{\sum n_y \cdot (y - y_{ср})^2 / n}$

Коэффициент корреляции принимает значения от –1 до+1.

С помощью получившихся коэффициентов корреляции можно составить корреляционную матрицу и построить графы связности признаков с учетом отброса статистически незначимых данных. Количество и толщина линий определяется рангом. Если ранг равен

нулю, то связи между признаками нет, если единице, то средняя, если ранг равен двум, то связь сильная, если трем, то очень сильная.

Чувствительность:

$$Se = PS / S \cdot 100\%, \qquad (1.9)$$

где PS – количество больных с идентифицированным значением признака, S – общее число больных.

Специфичность:

$$Sp = NH / H \cdot 100\%, \qquad (1.10)$$

где NH – число здоровых, у которых отсутствует рассматриваемый признак, H – общее число здоровых людей.

$$Эфф = \frac{PS + R}{S + H} \cdot 100\%, \qquad (1.11)$$

где R – число здоровых, попавших в доверительный интервал для здоровых, H - общее число здоровых людей, S – общее число больных.

Для построения решающих правил находим доверительный интервал по формуле: $P_0 \pm \Delta P$, где P_0 - среднее, $\Delta P = \delta_p / \sqrt{n-1}$.

Результаты исследования

Исходные данные представлены в таблицах 1.1 и 1.2.

Таблица 1.1 – Исходные данные

№	X1	X2	X3	X4	X5
1	2	3	4	5	6
1	211	-112	876	6981	1297
2	205	-100	873	6997	1297
3	208	-97	874	6998	1306
4	198	-85	889	6993	1293
5	199	-111	868	6995	

Продолжение таблицы 1.1

1	2	3	4	5	6
6	212	-95	876	6992	1301
7	199	-107	882	6993	1297
8	204	-102	875	6990	1295
9	198	-104	882	7001	1315
10		-103	877	7009	1293
11	207	-100	881	7010	1287
12	204	-104	867	7005	1307
13	221	-94	885	6984	1292
14	208	-101	856	7000	1307
15	189	-92	891	6994	1292
16	210	-105	898	7002	1293
17	25	-96	852	6998	1305
18	206	-92	877	6995	1315
19	193	-92	867	6995	1290
20	186	-97	883	6990	1309

Таблица 1.2 – Исходные данные

№	X11	X22	X33	X44	X55
1	2	3	4	5	6
1	168	-99	832	7017	1278
2	184	-103	910	7019	1335
3	169	-57	826	6996	1262
4	206	-68	857	6965	1296
5	261	-128	847	7025	1345
6	188	-69	889	7029	1308
7	156	-28	850	7046	1352
8	192	-107	925	7060	1266
9	241	-140	900	6963	1278
10	185	-90	857	7022	1336
11	139	-148	894	7023	1310
12	222	-63	837	6973	1255
13		-125	900	7034	1226
14	201	-118	921	7022	1355
15	222	-118	878	6975	1341
16	191	-99	836	6974	1336

Продолжение таблицы 1.2

1	2	3	4	5	6
17	239	-102	851	6975	1300
18	198	-139	847	7066	1239
19	179	-52	898	6945	1274
20	238	-115	849	7013	1320

Расчетная часть

Заполним пропуски. После заполнения получились таблицы 1.3 и 1.4.

Таблица 1.3 – Исходные данные с заполненными пропусками

№	X1	X2	X3	X4	X5
1	211	-112	876	6981	1297
2	205	-100	873	6997	1297
3	208	-97	874	6998	1306
4	198	-85	889	6993	1293
5	199	-111	868	6995	**1315**
6	212	-95	876	6992	1301
7	199	-107	882	6993	1297
8	204	-102	875	6990	1295
9	198	-104	882	7001	1315
10	**206**	-103	877	7009	1293
11	207	-100	881	7010	1287
12	204	-104	867	7005	1307
13	221	-94	885	6984	1292
14	208	-101	856	7000	1307
15	189	-92	891	6994	1292
16	210	-105	898	7002	1293
17	25	-96	852	6998	1305
18	206	-92	877	6995	1315
19	193	-92	867	6995	1290
20	186	-97	883	6990	1309

Таблица 1.4 – Исходные данные с заполненными пропусками

№	X11	X22	X33	X44	X55
1	2	3	4	5	6
1	168	-99	832	7017	1278
2	184	-103	910	7019	1335
3	169	-57	826	6996	1262
4	206	-68	857	6965	1296
5	261	-128	847	7025	1345
6	188	-69	889	7029	1308
7	156	-28	850	7046	1352
8	192	-107	925	7060	1266
9	241	-140	900	6963	1278
10	185	-90	857	7022	1336
11	139	-148	894	7023	1310
12	222	-63	837	6973	1255
13	**222**	-125	900	7034	1226
14	201	-118	921	7022	1355
15	222	-118	878	6975	1341
16	191	-99	836	6974	1336
17	239	-102	851	6975	1300
18	198	-139	847	7066	1239
19	179	-52	898	6945	1274
20	238	-115	849	7013	1320

Удалим артефакты. Все значения каждого параметра должны находиться в диапазоне M±2σ. Если это не так, то заменяем это значение медианой соответствующего показателя.

Найдем статистические показатели, необходимые для этого (таблицы 1.5-1.6).

Таблица 1.5 – Статистические показатели

Показатель	X1	X2	X3	X4	X5
Мат. ожидания	198,5	-104,5	879,5	6985,5	1303
Дисперсия	1591,25	70,05	125,25	163,05	82,8
СКО	39,89048	8,369588	11,19151	12,7691	9,099451

Таблица 1.6 – Статистические показатели

Показатель	X11	X22	X33	X44	X55
Мат. ожидания	203	-107	840,5	7015	1299
Дисперсия	1591,25	70,05	125,25	163,05	82,8
СКО	39,89048	8,369588	11,19151	12,7691	9,099451

В данном случае все значения попадают в данный диапазон. Таким образом, артефактов нет.

Найдем основные статистические характеристики данных показателей (таблицы 1.7-1.8).

Таблица 1.7 – Основные статистические характеристики данных показателей

Показатель	X1	X2	X3	X4	X5
Медиана	206,5	-101,5	879	7009,5	1290
Мода					
Эксцесс	18,15675	-0,18185	0,574618	0,26748	-1,10268
Коэффициент асимметрии	1,56607	982744	13,83187	4575619	2936,744
Коэффициент вариации	20,09596	-8,00917	1,272486	0,182794	0,698346
Минимум	25	-112	852	6981	1287
Максимум	221	-85	889	7010	1315

Таблица 1.8 – Основные статистические характеристики данных показателей

Показатель	X11	X22	X33	X44	X55
Медиана	162	-119	875,5	7022,5	1323
Мода					
Эксцесс	-0,43571	-0,51893	-1,34469	-0,93136	-1,06646
Коэффициент асимметрии	1,56607	982744	13,83187	4575619	2936,744
Коэффициент вариации	19,65048	-7,82204	1,331531	0,182026	0,700497
Минимум	139	-148	826	6945	1226
Максимум	261	-28	925	7066	1352

Если коэффициент As<0, то это левосторонняя асимметрия, если As>0, то правосторонняя. Коэффициент вариации служит для сравнения величин рассеяния по отношению к выборочной средней двух вариационных рядов.

Чем больше коэффициент, тем больше рассеяние.

Построим одномерные гистограммы для таблиц 1.1 и 1.2 соответственно, шаг в которых рассчитаем по формуле:

$$\text{Шаг} = (\text{максимальное значение} - \text{минимальное значение})/(1+\log_2 \text{N})$$

где в данном случае N-20.

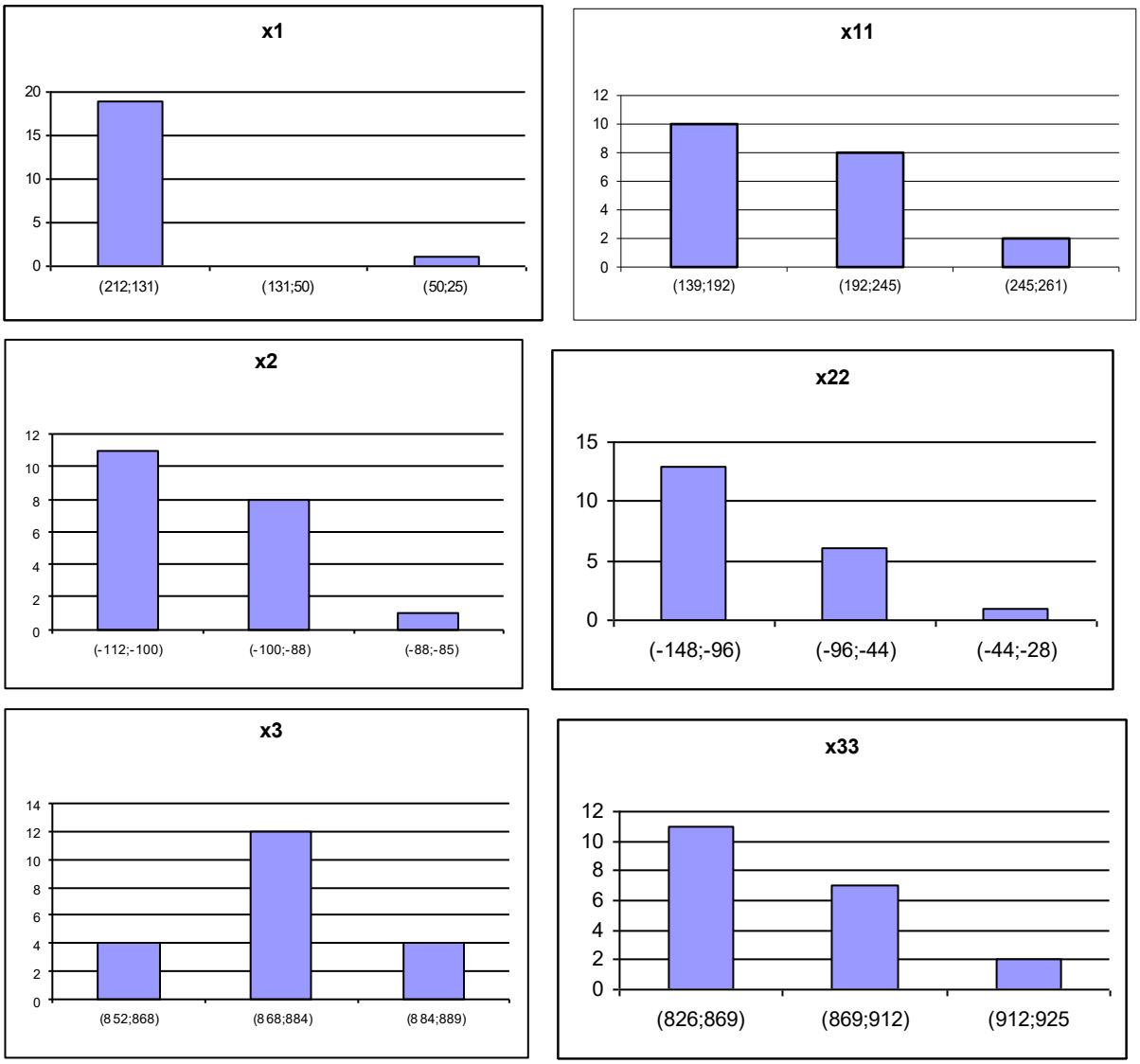


Рисунок 1.1 – Одномерные гистограммы для таблиц 1.1 и 1.2

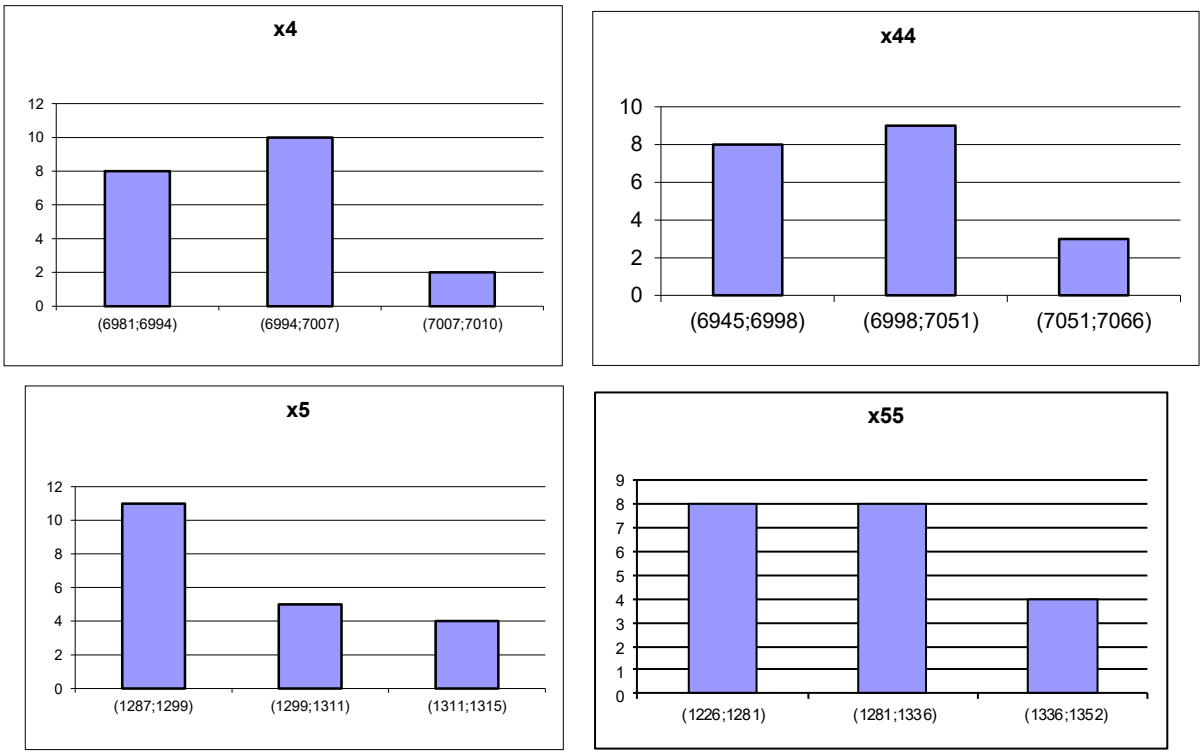


Рисунок 1.2 – Одномерные гистограммы для таблиц 1.1 и 1.2

Составим корреляционную матрицу. Для этого воспользуемся формулой 1.7.

Корреляционная матрица для таблицы 1.1 представлена в виде таблице 1.9.

Таблица 1.9 – Корреляционная матрица для таблицы 1.1

№	1	2	3	4	5
1	1	-0,1	0,5	-0,1	-0,1
2	-0,1	1	-0,01	0,46	-0,3
3	0,5	-0,01	1	-0,29	-0,26
4	-0,1	0,46	-0,29	1	-0,24
5	-0,1	-0,3	-0,26	-0,24	1

Корреляционная матрица для таблицы 1.3 представлена в виде таблицы 1.10.

Определим значимость каждого коэффициента корреляции и для каждого значимого коэффициента определим соответствующие регрессионные модели.

Таблица 1.10 – Корреляционная матрица для таблицы 1.3

№	11	22	33	44	55
11	1	-0,36	-0,1	-0,2	-0,03
22	-0,36	1	-0,03	-0,28	0,05
33	-0,1	-0,03	1	-0,05	0,05
44	-0,2	-0,28	-0,05	1	-0,02
55	-0,032	0,05	0,05	-0,02	1

Определим значимость каждого коэффициента корреляции и для каждого значимого коэффициента определим соответствующие регрессионные модели.

Если $\text{arctgR} > t(p)/\sqrt{n-3}$, где $t(p) = 1.67$, n – длина выборки, то коэффициент значим.

Итак, в нашем случае значимыми являются коэффициенты:

$R11 = R22 = R33 = R44 = R55 = 1$

$R1111 = R2222 = R3333 = R4444 = R5555 = 1$

$R14 = R41 = 0.5$

$R25 = R52 = -0.3$

Зависимость показателей для таблицы 1.1 и 1.2 показана на рисунках 3.3-3.4.

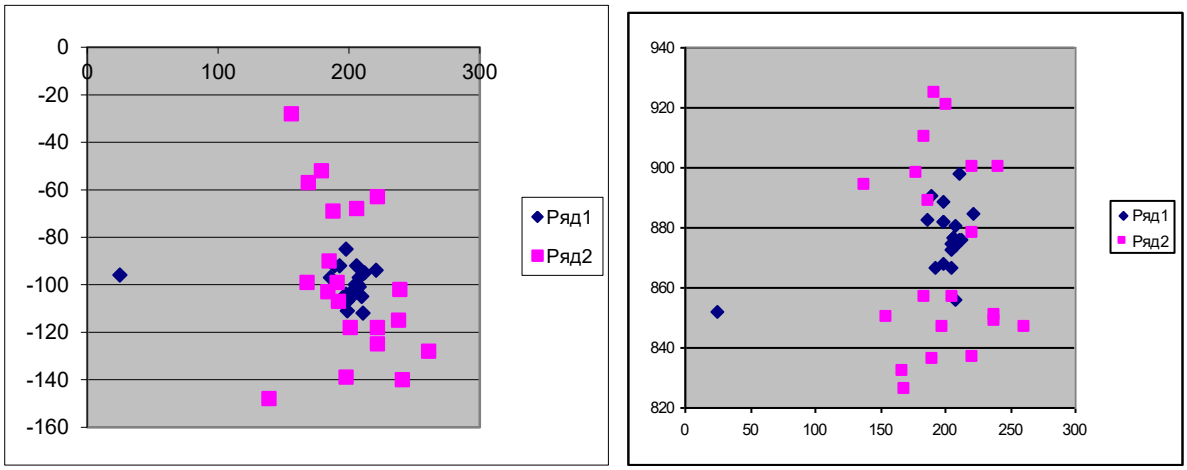


Рисунок 1.3 – Зависимость показателей для таблиц 1.1 и 1.2

По полученным данным найдем тип решающего правила.

Построим линейную разделяющую поверхность для двух классов.

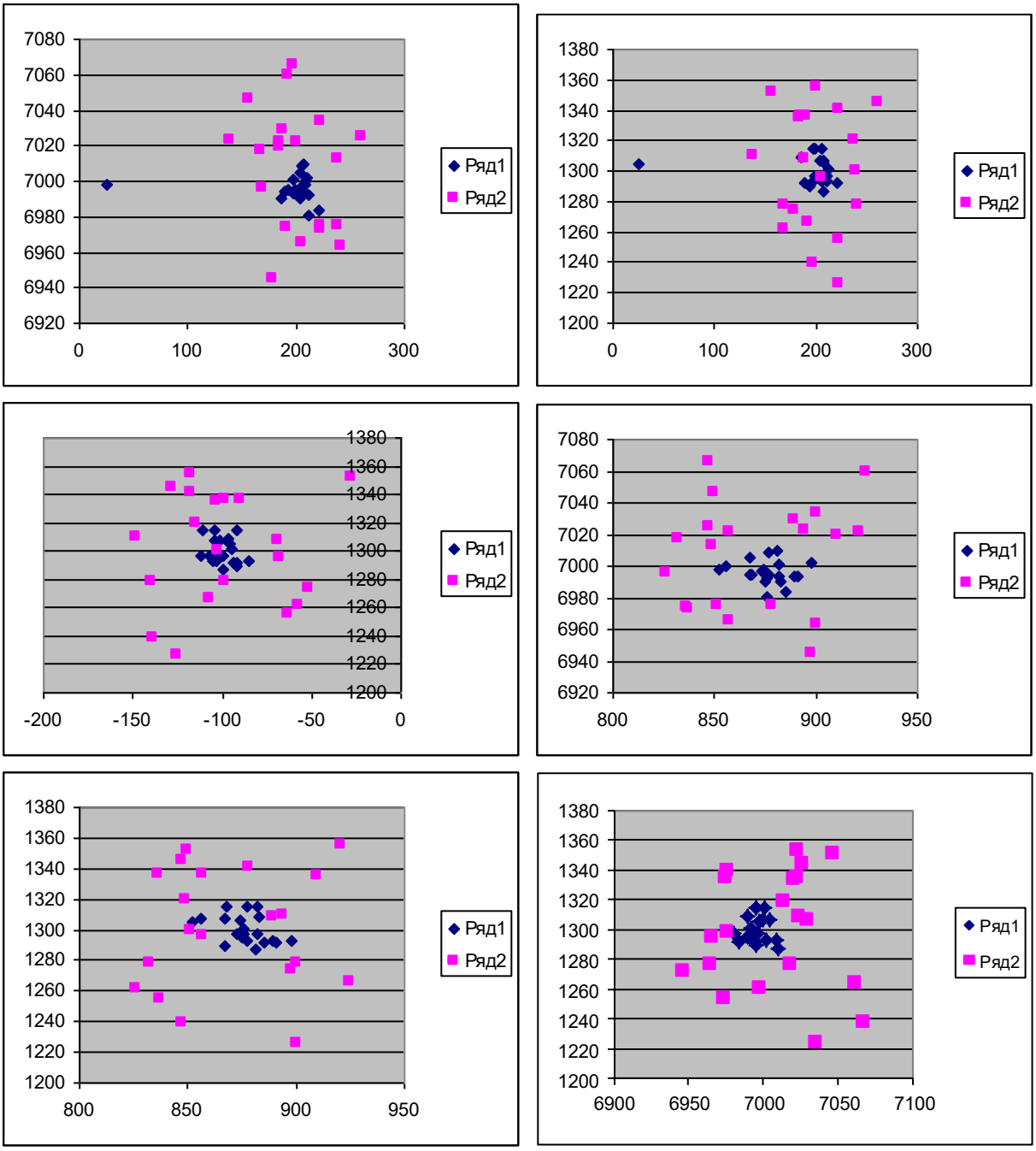


Рисунок 1.4 – Зависимость показателей для таблиц 1.1 и 1.2

ЛАБОРАТОРНАЯ РАБОТА №2. ПОСТРОЕНИЕ ЛИНЕЙНЫХ РАЗДЕЛЯЮЩИХ ПОВЕРХНОСТЕЙ

1. Порядок построения линейной разделяющей гиперплоскости

По полученным обучающим выборкам найдем разделяющую гиперплоскость, которая проходит через середину отрезка, соединяющего центроиды двух обучающих выборок и перпендикулярна к нему.

Построим линейную разделяющую поверхность для двух классов.

Если вектор X_i характеризует i-й объект первого класса, а вектор Y_j характеризует j–й объект второго класса, то координаты центроид A и B , а, следовательно, и отрезка AB, их соединяющего, определяются как M[X] и M[Y].

Координаты точки C, которая лежит на середине отрезка AB, определяются по формуле:

$$C((M[x_1] + M[y_1])/2; (M[x_2] + M[y_2])/2; ...; (M[x_N] + M[y_N])/2),$$

где N –число информативных признаков или размерность признакового пространства.

Если плоскость перпендикулярна вектору n(a; b; c), например, в трехмерном пространстве, то уравнение плоскости в этом пространстве записывается как:

$$ax + by + cz + d = 0.$$

Уравнение плоскости, перпендикулярной вектору n(a; b; c) и проходящей через точку (x0; y0; z0) записывается как:

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

Чтобы перейти от вектора AB к вектору **n**, необходимо из координат $M[x_i]$ вычесть координаты $M[y_i]$, то есть из координат точки A вычесть координаты точки B.

Для этого необходимо найти координаты точки C, которая характеризует среднюю составляющую по формуле:

$$C((M[x]1 + M[y]1)/2; (M[x]2 + M[y]2)/2; ...)$$

$$C(-25{,}25; 538{,}75; 2075; 3660{,}75; 175{,}75)$$

Теперь необходимо найти координаты вектора AB, которые соединяют точки X и Y.

$$AB(-11{,}5; -1242{,}5; 51{,}5; 4201{,}5; -30{,}5)$$

Найдем уравнение плоскости, которая проходит через точку C и имеет нормаль AB. Если принять, что AB(A, B,C,D) и $C(X_0,Y_0, Z_0, E_0)$, тогда плоскость будет иметь уравнение:

$$-11{,}5 \cdot (X1 + 25{,}5) - 1242{,}5 \cdot (X2 - 538{,}75) + 51{,}5 \cdot (X3 - 20{,}75) +$$
$$+ 4201{,}5 \cdot (X4 - 3660{,}75) - 30{,}5 \cdot (X5 - 175{,}75) = 0$$
$$-11{,}5 \cdot X1 - 1242{,}5 \cdot X2 + 51{,}5 \cdot X3 + 4201{,}5 \cdot X4 - 30{,}5 \cdot X5 - 18367996 = Y$$

Найдем значения Y для строк каждого класса и найдем минимум и максимум для них (таблицы 2.1-2.2)

Таблица 2.1 – Нахождение значения Y для строк каждого класса, минимума и максимума.

№	X1	X2	X3	X4	X5	y
1	2	3	4	5	6	7
1	211	-112	876	6981	1297	
2	205	-100	873	6997	1297	
3	208	-97	874	6998	1306	
4	198	-85	889	6993	1293	
5	199	-111	868	6995	**1315**	
6	212	-95	876	6992	1301	
7	199	-107	882	6993	1297	
8	204	-102	875	6990	1295	
9	198	-104	882	7001	1315	
10	**206**	-103	877	7009	1293	
11	207	-100	881	7010	1287	

Продолжение таблицы 2.1

1	2	3	4	5	6	7
12	204	-104	867	7005	1307	
13	221	-94	885	6984	1292	
14	208	-101	856	7000	1307	
15	189	-92	891	6994	1292	
16	210	-105	898	7002	1293	
17	25	-96	852	6998	1305	
18	206	-92	877	6995	1315	
19	193	-92	867	6995	1290	
20	186	-97	883	6990	1309	
max						
min						

Таблица 2.2 – Нахождение значения Y для строк каждого класса, минимума и максимума

№	X11	X22	X33	X44	X55	y
1	2	3	4	5	6	7
1	168	-99	832	7017	1278	
2	184	-103	910	7019	1335	
3	169	-57	826	6996	1262	
4	206	-68	857	6965	1296	
5	261	-128	847	7025	1345	
6	188	-69	889	7029	1308	
7	156	-28	850	7046	1352	
8	192	-107	925	7060	1266	
9	241	-140	900	6963	1278	
10	185	-90	857	7022	1336	
11	139	-148	894	7023	1310	
12	222	-63	837	6973	1255	
13	**222**	-125	900	7034	1226	
14	201	-118	921	7022	1355	
15	222	-118	878	6975	1341	
16	191	-99	836	6974	1336	
17	239	-102	851	6975	1300	
18	198	-139	847	7066	1239	
19	179	-52	898	6945	1274	

Продолжение таблицы 2.2

1	2	3	4	5	6	7
20	238	-115	849	7013	1320	
max						
min						

ЛАБОРАТОРНАЯ РАБОТА №3. ИЗУЧЕНИЕ ЭТАЛОННЫХ КЛАССИФИКАТОРОВ

1 Индивидуальные задания

Исходные данные для первой совокупности представлены в таблице 3.1.

Таблица 3.1 – Исходные данные для первой совокупности

№	Nобуч/N контрольной	Математические ожидания информативных признаков						Дисперсии информативных признаков					
1	30/20	-18	112	101	495	10	51	100	500	100	1500	10	80
2	40/35	-21	109	140	491	17	49	81	145	39	1600	9	81
3	35/40	-21	100	132	513	20	-49	144	98	45	1460	9	144
4	15/30	20	95	120	499	16	-40	64	60	78	1000	18	64
5	18/25	26	73	66	509	17	-46	49	100	87	1000	20	49
6	25/19	-19	91	107	122	17	66	100	98	90	140	25	100
7	34/20	-15	68	93	122	23	-43	36	45	65	90	25	36
8	41/35	-19	72	78	127	13	-49	49	40	66	60	16	49
9	36/42	16	148	72	123	87	42	9	36	78	100	100	9
10	18/30	17	70	118	128	86	-60	16	26	76	98	100	16
11	19/25	-18	83	97	115	82	52	25	28	78	45	64	25
12	28/19	-21	107	141	124	80	-58	25	56	98	40	64	25
13	30/23	-22	112	97	127	70	55	49	80	100	36	49	49

Исходные данные для второй совокупности представлены в таблице 3.2.

Таблица 3.2. – Исходные данные для первой совокупности

№	Nобуч/ Nконтрольной	Математические ожидания информативных признаков						Дисперсии информативных признаков					
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	20/40	-8	120	90	490	15	61	10	100	100	400	10	36
2	40/35	-25	209	150	441	27	69	8	145	39	500	9	49
3	42/40	-28	110	80	413	27	-29	14	98	45	300	9	36
4	19/30	30	85	128	409	19	-30	6	60	78	100	18	9
5	18/25	26	93	76	559	27	-56	4	100	87	100	20	49
6	35/19	-9	71	88	122	27	86	10	98	90	100	25	49
7	24/20	-5	70	89	132	33	-33	8	45	65	80	25	36
8	31/35	-9	92	60	147	23	-69	4	40	66	9	16	49
9	36/42	20	120	87	133	97	52	9	36	78	10	10	36
10	48/30	20	80	81	158	116	-50	16	26	76	9	10	16

Продолжение таблицы 3.2

11	19/25	-19	78	100	125	102	82	25	28	78	4	4	25
12	28/19	-29	120	131	134	89	-78	25	56	98	4	4	25
13	35/25	-28	140	87	147	80	59	4	80	100	6	9	49

2 Получение контрольных и обучающих выборок

Используя генератор случайных чисел пакета MathCad можно создать как обучающие, так и контрольные выборки с любым объемом и любой размерности пространства информативных признаков. Пример листа MathCad, в котором создается выборка из 28 объектов с нормально распределенным законом распределения признаков (признаковое пространство четырехмерное) показан на рисунок 3.1.

$n := 28$ $m := 4$ $i := 0..m-1$ $j := 0..n-1$

$s_0 := 225$ $sd_0 := 16$ $s_1 := 228$ $sd_1 := 18$ $s_2 := 231$

$sd_2 := 15$ $s_3 := 235$ $sd_3 := 16$

$$V\lambda_i := \overrightarrow{ceil\left(rnorm\left(n, s_i, sd_i\right)\right)}$$

Рисунок 3.1 – Лист MathCad с генератором случайных чисел

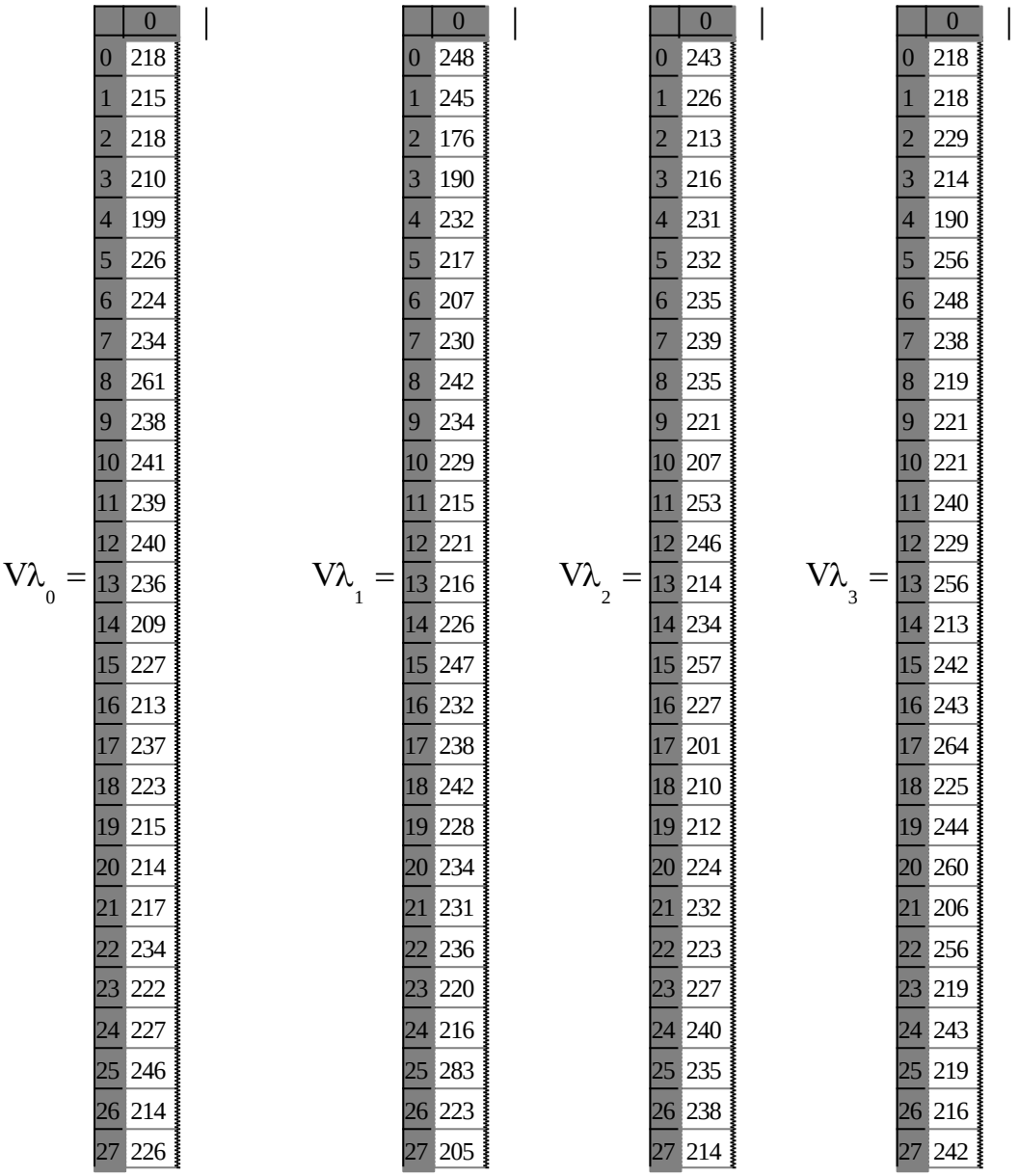
Каждый информативный признак в выборке представлен вектором-столбцом Vλ из 28 элементов. Случайные числа, распределенные по нормальному закону распределения, генерируются функцией rnorm(n, s, sd), где n-число элементов в выборке, s – математическое ожидание информативного признака, sd -дисперсия информативного признака.

Варьируя эти параметры (s и sd) вы можете менять структуру распределения классов в признаковом пространстве и исследовать эффективность классификации при различных классовых структурах.

Функция ceil округляет случайное число до ближайшего целого.

Значок → обозначает операцию векторизации, то есть одновременное выполнение скалярной операции над всеми элементами вектора.

Пример полученной выборки в четырехмерном признаковом пространстве показан на рисунке 3.2.



	0	Vλ0	Vλ1	Vλ2	Vλ3
0	218	248	243	218
1	215	245	226	218
2	218	176	213	229
3	210	190	216	214
4	199	232	231	190
5	226	217	232	256
6	224	207	235	248
7	234	230	239	238
8	261	242	235	219
9	238	234	221	221
10	241	229	207	221
11	239	215	253	240
12	240	221	246	229
13	236	216	214	256
14	209	226	234	213
15	227	247	257	242
16	213	232	227	243
17	237	238	201	264
18	223	242	210	225
19	215	228	212	244
20	214	234	224	260
21	217	231	232	206
22	234	236	223	256
23	222	220	227	219
24	227	216	240	243
25	246	283	235	219
26	214	223	238	216
27	226	205	214	242

Рисунок 3.2 – Пример выборки из 28 элементов в четырехмерном признаковом пространстве

По данным выборок строится нейронная сеть, содержащая по n нейронов в слое. Обучение производится в 50 циклов при предельном значении критерия обучения 0.07. Примеры

результатов обучения нейронной сети с помощью программы Neurowork приведен на рисунках 3.3-3.4.

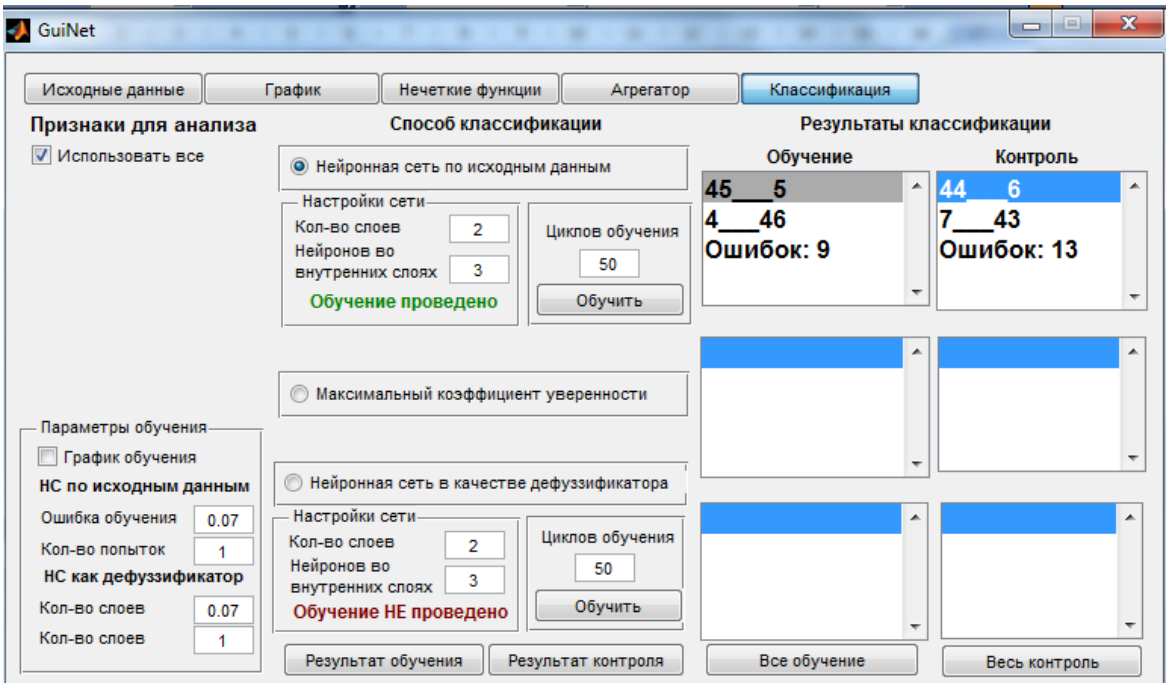


Рисунок 3.3 - Результаты обучения и контроля выборок двухслойной нейронной сети

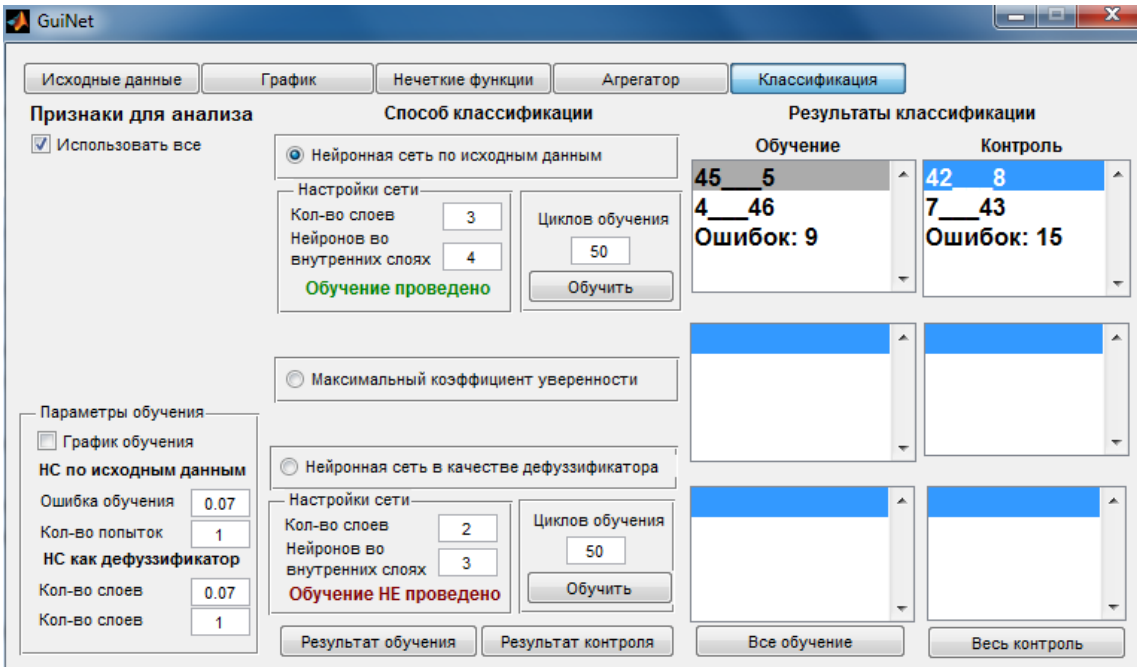


Рисунок 3.4 - Результаты обучения и контроля выборок трехслойной нейронной сети, содержащей по 4 слоя

3 Оценка эффективности методов распознавания

В качестве расчетных показателей качества диагностических решающих правил используется: диагностическая чувствительность (ДЧ), диагностическая специфичность (ДС), прогностическая значимость положительных результатов ($ПЗ^{+}$), прогностическая значимость отрицательных результатов ($ПЗ^{-}$), диагностическая эффективность решающего правила (ДЭ).

Эти показатели вычислялись по данным таблицы распределений результатов контрольных испытаний (таблица 3.3).

Таблица 3.3 – Таблица контрольных испытаний

Обследуемые	Результаты срабатывания правил		Всего
	положительные	отрицательные	
n_{ω_r}	ИП	ЛО	ИП+ЛО
n_{ω_0}	ЛП	ИО	ЛП+ИО
Всего	ИП+ЛП	ЛО+ИО	ИП+ЛП+ЛО+ИО

где r – номер класса исследуемого заболевания; n_{ω_r} - количество людей в контрольной выборке в исследуемом классе заболеваний; n_{ω_0} - количество здоровых людей в контрольной выборке; ИП – истинно положительный результат равный количеству людей класса ω_r правильно классифицируемых рассматриваемым правилом; ЛП – ложно положительный результат равный количеству людей класса ω_0 ошибочно отнесенных решающим правилом к классу ω_r; ЛО – ложно отрицательный результат: количество людей класса ω_r ошибочно отнесенных решающим правилом к классу ω_0; ИО – истинно отрицательный результат: количество людей класса ω_0 правильно классифицируемых решающим правилом.

Для приведенных в таблице 3.3 обозначений расчет показателей качества осуществляется в соответствии с выражениями:

$$\begin{cases} ДЧ = ИП / n_{\omega_г} \\ ДС = ИО / n_{\omega_0} \\ ПЗ^{+} = ИП / (ИП + ЛП) \\ ПЗ^{-} = ИО / (ЛО + ИО) \\ ДЭ = (ИП + ИО) / (ИП + ЛП + ЛО + ИО) \end{cases}$$

Список литературы

1. Боровиков, В. STISTICA. Искусство анализа данных на компьютере: Для профессионалов [Текст] / В. Боровиков. 2-е изд. (+CD). СПб.: Питер, 2003. 688 с.
2. Горелик, А.Л. Методы распознавания: Учеб. пособие для вузов [Текст] / А.Л. Горелик, В.А. Скрипкин. М.: Высшая школа, 2004. 261 с.
3. Омельченко, В.П. Практикум по медицинской информатике [Текст] : серия «Учебники, учебные пособия» / В.П. Омельченко, А.Л. Демидова. Ростов-на-Дону: Феникс, 2001. 304 с.